Article

A topological look into the evolution of developmental programs

Somya Mani[1,*] and Tsvi Tlusty[1,2,*]

[1]Center for Soft and Living Matter, Institute for Basic Science, Ulsan, Republic of Korea and [2]Departments of Physics and Chemistry, UNIST, Ulsan, Republic of Korea

ABSTRACT   Rapid advance of experimental techniques provides an unprecedented in-depth view into complex developmental processes. Still, little is known on how the complexity of multicellular organisms evolved by elaborating developmental programs and inventing new cell types. A hurdle to understanding developmental evolution is the difficulty of even describing the intertwined network of spatiotemporal processes underlying the development of complex multicellular organisms. Nonetheless, an overview of developmental trajectories can be obtained from cell type lineage maps. Here, we propose that these lineage maps can also reveal how developmental programs evolve: the modes of evolving new cell types in an organism should be visible in its developmental trajectories and therefore in the geometry of its cell type lineage map. This idea is demonstrated using a parsimonious generative model of developmental programs, which allows us to reliably survey the universe of all possible programs and examine their topological features. We find that, contrary to belief, tree-like lineage maps are rare, and lineage maps of complex multicellular organisms are likely to be directed acyclic graphs in which multiple developmental routes can converge on the same cell type. Although cell type evolution prescribes what developmental programs come into existence, natural selection prunes those programs that produce low-functioning organisms. Our model indicates that additionally, lineage map topologies are correlated with such a functional property: the ability of organisms to regenerate.

SIGNIFICANCE   Cell type invention is a chief process in the evolution of developmental programs. Traditionally, developmental trajectories are represented as cell type lineage maps. Here, we propose that systematic analysis of these maps, in particular their topology, should reveal traces of the manner in which cell types were invented. This is illustrated using a generative model of developmental programs, which allows one to robustly survey the geometry of cell-lineage maps and link them to modes of cell type invention. We suggest that predictions made by such mathematical models, in conjunction with surveys of real cell-lineage maps of different multicellular lineages, could uncover mechanisms underlying evolution of developmental programs.

STATISTICS OF LINEAGE MAPS REFLECT DEVELOPMENTAL EVOLUTION

How can one understand the astounding richness of life forms? Although molecules and mechanisms of biological development are conserved within each multicellular lineage (1), these lineages are extraordinarily diverse; land plants and animals include many thousands to millions of species (2). This diversity is in part due to the distinct cell types present in different organisms. And in this sense, (Appendix) developmental programs evolve by inventing new cell types (3). Although ancestral lineages likely resem-

bled the simplest multicellular organisms alive today, such as *Volvox carterii*, which has two cell types (4), the extant diversity of today's organisms ranges from those with a few cell types to those with hundreds. What molecular mechanisms and logic could produce such diversity remains a persistent question in development.

One way to tackle this question is by comparing developmental programs across species of various levels of (Appendix) complexity. Analyzing developmental genes to see how gene families have expanded is fairly easy. But we now realize that this is far from sufficient because genes interact combinatorially to express cell types. For example, looking at genomes of sponges, one might be tempted to conclude that they possess neurons because they have all the necessary components to make synapses. In reality, however, it was only in the bilaterian/cnidarian ancestor that these

components were arranged in a manner that expresses the synapse ([3](#)).

This example is a reminder that developmental programs are functions or algorithms for the assembly of organisms. Long ago, it was recognized by Cantor that there are always many more conceivable functions than combinations of variables ([5](#)). Perhaps the simplest example—which is used as a minimalist model of development ([6](#))—is a Boolean function of N binary variables. There are 2^N combinations of variables, but a much larger number, 2^{2^N}, of possible Boolean functions. For finite sets, such as the N binary variables, the result is proved very simply by enumerating all possible functions. The groundbreaking discovery of Cantor was that the result applies also to infinite sets. Cantor's theorem has two implications relevant to our discussion:

1) First, one cannot infer a function merely by inspecting its variables. Or equivalently, listing the lines of a computer program without knowing how they are logically linked will not suffice to understand the algorithm. Thus, to describe development in any organism, it is essential to look at developmental genes in the context of their regulatory architecture.

2) However, as the complexity of the organism increases, this task quickly becomes infeasible because the number of possible functions or programs increases very steeply with the number components. This number grows super-exponentially, like the number 2^{2^N} of Boolean functions. For example, $N = 10$ binary variables have $2^{10} = 1024$ combinations, whereas the number of Boolean functions of 10 variables is $2^{2^{10}} \simeq 10^{308}$, much more than the number of particles in the universe. Thus, for very simple organisms, such as *Volvox* with its two cell types, one may outline a complete picture of its developmental process ([4](#)). But in complex organisms, such as humans with their 200 cell types ([7](#)), development is an elaborate process that involves coordinated gene expression, cell-cell communication, asymmetric cell division, cell movements, and cell death ([8](#)).

Such combinatorially exploding complexity is daunting. Still, we do have a succinct, accessible representation of developmental events in terms of cell type lineage maps (CTLMs ([Appendix](#))) that detail how different cell types of the body are generated through cellular differentiation. Besides cataloguing differentiation events, CTLMs provide a glimpse into the underlying regulatory architecture. Implicit in these maps is information about which cell states are stable, and thus can be called cell types, into how many cell types any given type can differentiate and how differentiation depends on ([Appendix](#)) cellular context.

We propose here that CTLMs may also teach us about the evolution of developmental programs. Invention of a new cell type involves rewiring of the underlying regulatory architecture of development. This architecture controls not only the identities of cell types but also the developmental

trajectories through which they are produced. Therefore, it is reasonable to assume that the mechanisms that lead to cell type invention should leave a mark on developmental trajectories and on the geometry of lineage maps that essentially trace these trajectories. Indeed, in ([9](#)), through a comparison of cell-lineage maps of two nematode species, the authors were able to identify evolutionary correspondence between their cell types. Here, we further suggest that the statistics of the ([Appendix](#)) topology of lineage maps should reflect the modes through which cell types evolve.

In this article, we first describe, to the best of our knowledge, the role CTLMs play in the study of biological development and dwell on prevalent biases in our conception of CTLMs. We then consider the role CTLMs could play in elucidating developmental evolution and discuss this idea using a simple generative model ([10](#)). The model allows us to anticipate the statistical distribution of lineage graphs generated by distinct modes of cell type evolution. We examine a simple case ("null hypothesis") in which developmental programs are not biased by any particular mode of cell type invention. Contrary to ingrained belief, our model shows that CTLMs are highly unlikely to be tree-like. Instead, they are likely to be directed acyclic graphs (DAGs) in which a single-cell type is reachable via multiple developmental routes.

The model also demonstrates how topologies of CTLMs can encode information about functional attributes of organisms. For instance, we see that the ability of organisms to regenerate is correlated with lineage map topology. Such correlations can form a basis for natural selection to favor certain developmental programs and thereby reveal more layers in the multilevel process of developmental evolution.

USEFULNESS OF CTLMs IN THE STUDY OF DEVELOPMENT

Bodies of multicellular organisms are composed of multiple types of cells, differing by the distinct functions they perform. For example, humans with their ≥ 200 cell types ([7](#)) are composed of cells such as neurons that process and relay information, muscle cells that allow locomotion, B cells that provide immunity, etc.

Irrespective of its complexity, any adult multicellular body is ultimately derived from a single-celled zygote through the process of development. Historically, to trace development, embryonic cells were labeled with dyes that allowed following their divisions to identify the adult tissues they form. This mapping of embryonic cells to cell types in the adult is called fate mapping. The construction of fate maps provided insights into the mechanism of fate determination during development. Certain embryonic cells are autonomously specified, meaning that their fates remain unchanged even when they are grown separately. For example, in the tunicate *Styela partitia*, any separated embryonic cell type gives rise to the same adult cell types it normally does

and therefore yields partial adults. Other embryonic cells are conditionally specified, for example in sea urchins, in which separated embryonic cells behave differently and develop into complete adults (8).

Such mappings of cells produced by following sequential cell divisions are still useful today as a means to understand and represent developmental mechanisms. For example, in female *Drosophila*, the cell divisions that give rise to the oocyte, and associated nurse cells are highly regular and yield a characteristic lineage tree. The closeness of nurse cells to oocytes in this lineage tree strongly determines their cell size (11).

Molecular markers allow identification of cell types much more accurately and thereby facilitate refinement of fate maps. The most refined versions of fate maps are called CTLMs, and these track every single cellular differentiation event between the initial embryonic cells and the final adult cells. Mathematically speaking, a CTLM is a graph in which the nodes represent distinct cell types, and directed edges represent differentiation of one cell type into another (Fig. 1).

CTLMs capture progression through life stages

Generally, CTLMs can represent not only embryonic development but other life stages as well, such as adult homeostasis (14), metamorphosis (15), and regeneration (16). In reality, especially for asexually reproducing organisms, it is not always possible to tell apart the CTLMs of different life stages. For example, Fig. 1 *C* can be said to represent both the embryonic development and the adult homeostatic map for *Volvox* (4), and although Fig. 1 *D* represents the CTLM for adult homeostasis in hydra (13), it contains edges representing cellular differentiation that are also observed during hydra regeneration (16). That is, these organisms reuse the same differentiation pathways, and a single

CTLM can sufficiently describe various life stages. In contrast, developmental programs can also display extreme plasticity; for example, under unfavorable conditions, the immortal jellyfish *Turritopsis dohrnii* can reverse its development (17), essentially reversing edges across its CTLM.

Fundamentally, CTLMs of various life stages of an organism reflect the parts of its "regulatory architecture"—the intercellular signaling system and the gene regulatory network (GRN)—that are accessed during these different stages in the organism's lifetime. And the plasticity of CTLMs indicates that multiple developmental routes can potentially be used to access the same cell type. Which developmental route is eventually realized depends on the succession of cellular contexts a cell type encounters during its differentiation.

Tracing CTLMs with single-cell transcriptomics

The first step in the construction of CTLMs is obviously the identification of an organism's different cell types. Some cell types can be identified simply by their distinctive morphological features, such as the axonal projections of neurons, the striped appearance of striated muscle cells, the disk-like shape of erythrocytes, etc. But such morphological and functional descriptions might be misleading. For example, smooth muscles in vertebrates and striated muscles in *Drosophila* are morphologically and functionally distinct but evolutionarily and developmentally equivalent (18). Hence, how these different cells gain the ability to perform their functions and how they are related to other cell types can only be seen through their molecular-level descriptions and gene expression patterns (3).

Recent advances in techniques such as single-cell transcriptomics allow us to identify cell types with unprecedented accuracy and detail. These catalogs of cell types on their own already produce insights into the functioning
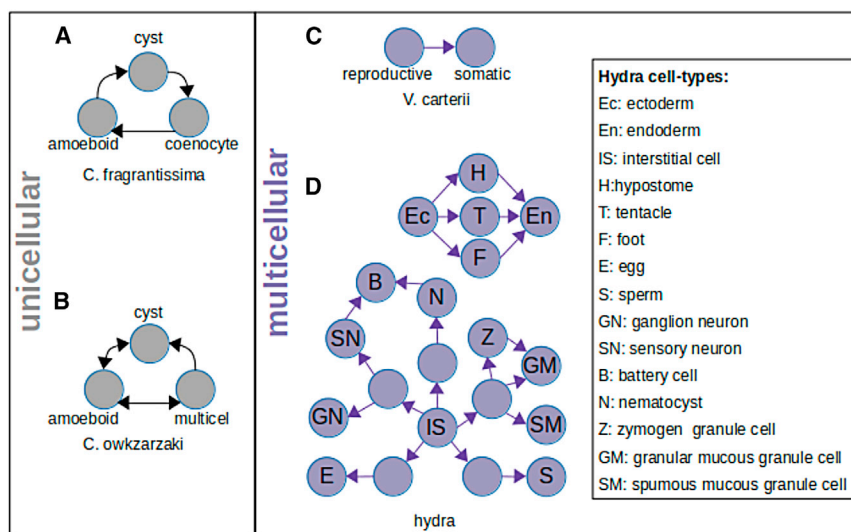

FIGURE 1 Cell type lineage maps (CTLMs). In (*A*) and (*B*), gray circles represent cell states of unicellular organisms, and arrows represent changes in cellular phenotype. In (*C*) and (*D*), the blue circles represent cell types of multicellular organisms, and arrows represent differentiation. (*A*) The life cycle of the unicellular organism *Creolimax fragrantissima* involves cycling between three stages: a motile amoeboid state, the immobile cyst state, and the multinucleate coenocyte (12). (*B*) *Capsaspora owkzarzaki* responds to environmental cues to switch between three cell-states: cells switch from a reproductive amoeboid state to an aggregative multicellular state in the presence of nutrients, and both the amoeboid and multicellular cell states switch to a cyst state under starvation (12). (*C*) The CTLM for both embryonic development and adult homeostasis of *V. carterii* (4). (*D*) The CTLM representing adult homeostasis in hydra (13). Empty circles represent unannotated intermediate cell types. To see the figure in color, go online.

of an organism; for example, the presence of neurons indicates that the organism is capable of transmitting information across its body. But to gain insights into the process of development, cell types need to be arranged into CTLMs (19).

Several recent studies already report CTLMs reconstructed from single-cell transcriptomics data (13,14). However, there still remain many technical challenges in interpreting these data (20). Most notably, the algorithmic methods used to infer lineage maps are biased to preferentially produce (Appendix) tree-like and (Appendix) chain-like topologies (19).

The deeply rooted idea that CTLMs resemble trees probably owes itself to the history of cell-lineage maps: the first cell-lineage map of embryonic development ever constructed was that of *Caenorhabditis elegans*, and this map is remarkably tree-like (21). Additionally, another extensively studied lineage map is that of human hematopoiesis, which is also tree-like (22). Perhaps the "tree archetype" has persisted because of the tendency of the human mind to extrapolate: cells of multicellular organisms typically divide by binary fission. Evidently then, the two resulting daughter cells may assume at most two distinct cell fates. In addition, experimental studies of differentiation typically look at the conditions that lead a daughter of a stem cell toward one of two alternative cell fates (23,24). The inverse problem of how two different cell types can differentiate convergently to produce the same cell type is much less studied.

Although it is true that cellular differentiation is a branching process, it is erroneous to conclude that development, by extension, must look like a binary tree. In the first place, branches are parts of all connected graphs excluding simple chains and elementary cycles. Moreover, apart from the famous examples of tree-like lineage maps, we now also have examples of nontree-like lineage maps: zebrafish development (25) and hydra adult homeostasis (13).

As an aside, an interesting recent theoretical work looks at how tree-like spatial arrangement of cells could facilitate the origin of multicellularity through the specialization of reproductive germ cells and nonreproductive somatic cells (26). But even if we assume that the spatial arrangement of cells imposes constraints on the differentiation of incipient multicellular species, it does not exclude the possibility that further elaboration of this species could lead to nontree-like differentiation trajectories.

For all these reasons, it is essential that we deal with the biases in lineage map reconstruction procedures so that the resulting picture of development remains faithful to biological reality. Some studies overcome these biases by using single-cell transcriptomics alongside cellular barcoding (25,27), which allows unambiguous identification of cellular lineages (28). It is instructive to note the progress made in phylogenetics, another field in which the idea of phylogenetic "trees" is prevalent. Recent phylogenetic inference algorithms allow the inclusion of reticulate events

like hybridization and horizontal gene transfer (29), which yield nontree-like networks. These methods could illustrate how one could resolve biases toward tree-like topologies in cell-lineage reconstruction algorithms.

Elaboration of organisms through cell type invention

Traditionally, cell types are a concept relating to multicellular organisms, but unicellular organisms are also known to switch between cellular phenotypes. They do so either according to temporal programs (Fig. 1 *A*) or in response to changes in their environment (Fig. 1 *B*) (12). Cells in yeast colonies even show spatial organization of distinct cellular phenotypes (30). Moreover, the unicellular organisms that are closest to multicellular lineages possess homologs of many "multicellularity related" genes; for example *Chlamydomonas reinhardtii*, the closest unicellular relative of *Volvox*, has a homolog of regA, the gene responsible for cell type differentiation in *Volvox* (31).

A plausible origin of multicellularity was a transformation from an environmental or temporal regulation of cellular phenotypes to a developmental or spatial one. For example, in *Volvox*, regA expression causes differentiation of the reproductive cell to the somatic cell (Fig. 1 *C*), whereas the homolog of regA in *C. reinhardtii* responds to environmental stresses and induces a similar switch from a reproductive to a nonreproductive cellular phenotype (31). A similar scenario played out in the case of the facultatively multicellular dictyostid amoebas; the signaling molecules responsible for differentiation in multicellular species are instead produced in response to cold stress in unicellular dictyostelids (32).

Multicellularity has evolved multiple times in the history of life, but multicellular organisms have expanded their cell types and evolved complexity only in six lineages: once in animals, once in plants, twice in algae, and twice in fungi (2). The mechanism underlying the origin of multicellularity, which involved internalization of environmental cues, also plays a role in the invention of new cell types in already multicellular organisms. For example, the invention of the DSC cell type in placental mammals involved the internalization of stress signals into a cue for the differentiation of the ESF cell type. It also required rewiring of gene regulation such that the stressed ESF cell state, which would normally relax back to its nonstressed state, now differentiates into a new cell type, the DSC (Fig. 2).

Cell type evolution encoded in gene regulation and reflected in developmental programs

In (3), the authors anticipate changes in the genetic architecture that could lead to the invention of new cell types. To do this, they first emphasize the distinction between the function a cell type assumes in an organism and the gene
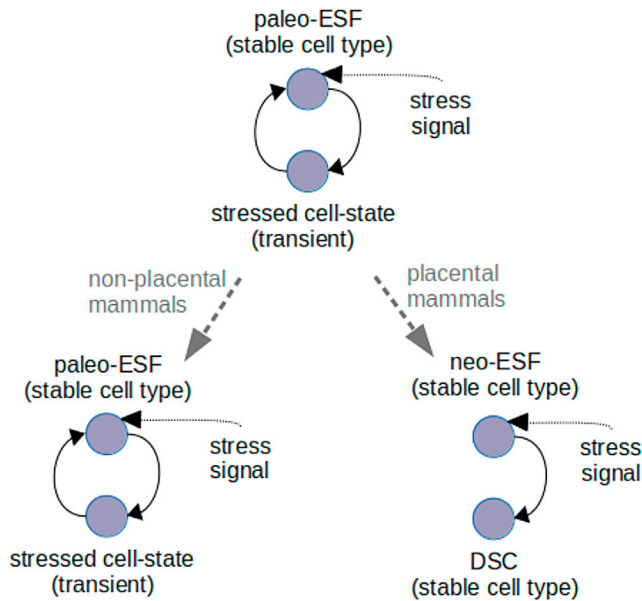
FIGURE 2 Evolution of GRN in mammals leads to the invention of a new cell type. Blue circles represent cell states, and bold arrows represent a change in cell state. Dashed gray arrows represent evolutionary transitions. Dotted black arrows represent the application of stress signal. In the ancestor of marsupials (nonplacental) and placental mammals, paleo-ESF cells responded to stress signals by elevating the expression of genes associated with stress responses and apoptosis. The stressed cell state then relaxes back into the normal paleo-ESF cell. But in placental mammals, a rewiring of the regulatory network led to the invention of two new cell types: the neo-ESF and the DSC cells. The neo-ESF cell, upon receiving stress signals, differentiates into the DSC cell instead of expressing stress response genes (33). To see the figure in color, go online.

regulation that ensures its stability. In other words, cell types are stable states of (Appendix) GRNs (see (34) for a treatment of GRNs as random Boolean networks).

Now, a multicellular organism can only contain a subset of the stable states prescribed by its GRN: these are the cell types that are accessed by its developmental program. GRNs also specify the dynamics through which transient cell states transform into stable cell types (35). We can imagine a range of molecular mechanisms—such as mutations affecting the activity of gene products or altering molecular interactions, etc.—that could rewire the GRN. And there are three ways in which such rewiring can lead to cell type "invention": 1) a frequently encountered transient cell state shifts the stable cell type it maps to (this mode probably led to the invention of DSC cell type; see Fig. 2); 2) a transient cell state becomes stable; and 3) the GRN expands by adding new genes, such as transcription factors, through gene duplication and divergence, horizontal transfer, etc. (2).

The prevalent modes of cell type invention could be different in different multicellular lineages. We expect that developmental trajectories, and therefore CTLMs, reflect these frequently occurring modes. At the same time, we are aware that the phylogenetic trajectories of cell types

and developmental trajectories need not be isomorphic (3). That is, although we expect modes of cell type invention to impose constraints on the topology of CTLMs, the exact form of this constraint is not obvious.

Moreover, development is also plastic, and trajectories leading to particular cell types can shift. For example, in the ancestral multicellular dictyostelids, stalk cells likely differentiated from prespore cells, whereas in the more recent group 4 dictyostelids, stalk cells differentiate from the developmentally distinct prestalk cells (32). The potential for this plasticity can also be seen in our own cells, in which expression of a single transcription factor, MYOD, can switch a fibroblast into a skeletal muscle cell (3). Over time, this plasticity can reconfigure CTLMs and potentially erase the signatures of the history of cell type invention.

From all this, we see that the question of developmental evolution can be broken down into two parts: 1) the manner in which evolutionary trajectories of cell types shape developmental trajectories, and 2) the extent to which this signature is preserved despite the plasticity of development. In (9), the authors demonstrate the relationship between cell type evolution and CTLM topology; they show that the evolutionary closeness of cell types can be identified through a comparison of CTLM subgraphs rooted at these cell types. We suggest here that beyond the detection of evolutionary closeness, global patterns of cell type evolution within multicellular lineages can be revealed through statistical analyses of CTLM topologies.

ANTICIPATING MODES OF CELL TYPE INVENTION FROM A GENERATIVE MODEL

As a concrete example for how mathematical models can bridge the regulatory architecture of development and CTLM topologies, we examine here a minimal generative model (10). Our model incorporates three fundamental features of development:

1) There exist multiple stable cell types and transient cell states that reliably map to specific stable cell types.
2) The fate of any cell depends on its cellular context due to cell-cell signaling. For example, during *Drosophila* oogenesis, cellular exchange of growth regulating proteins among the oocyte and surrounding nurse cells sets up a spatial coordinate system that determines the subsequent growth behavior of nurse cells (36).
3) Because development of a multicellular organism begins with a single cell, mechanisms for symmetry breaking, such as asymmetric cell division and cell polarization, that can create an interactive field of cells, are essential.

A generative model of development

We capture these features of development by decomposing the regulatory architecture of development into three

universal components: asymmetric cell division, cell signaling, and gene regulation. For simplicity, we encode these three components as Boolean logical functions. This coarse-grained representation does not depend on details specific to any particular organism and can therefore describe a wide variety of organisms. The simplicity allows us to sample millions of developmental programs as combinations of rules for cell division, signaling, and gene regulation (Fig. 3).

In the model, "genes" of organisms encode for cell-state determinants, which are interacting sets of transcription factors that combinatorially determine the identity of the cell (3). For an organism with N genes, cell states are defined as binary strings, in which "1" denotes the presence and "0" denotes the absence of a cell-state determinant. Cells in the model divide asymmetrically to produce unequal daughter cells (Fig. 3 A). Of the 2^N possible cell states, a few are assigned the status of stable cell types, whereas other cell states are transient and map deterministically to one of the stable cell types (Fig. 3 C). These assignments represent gene regulation in the model. Finally, certain cell-state determinants can be exchanged as signals among specific donor and receiver cells (Fig. 3 B). This allows cells to interact with their cellular context.

We model development as an unfolding process, which begins with a randomly picked initial cell type, the zygote, and proceeds through rounds of asymmetric cell division, cell-cell signaling, and gene regulation (Fig. 3 D). This sequence of operations is repeated until the resulting set of cell types repeats itself. This final set of cell types is a steady state of the model, which we call the adult. Within an organism in the model, the regulatory architecture is fixed throughout the development process. This algorithm can also model regeneration; here, instead of starting with the zygote, we represent an injury as a loss of a subset of adult cell types and initialize the developmental program with the remaining adult cell types.

In the model, a cell type x differentiates into a cell type y if one of the daughter cells of x gives rise to y after one round of signaling and gene regulation. In this way, we can construct CTLMs, which are graphs in which the nodes represent cell types, and directed edges represent differentiation (Fig. 3 E). A CTLM generated by the model can represent various stages of development:

1) Embryonic development when the nodes of the graph include all the cell types encountered, starting from the initial zygotic cell to the final adult.
2) Adult homeostasis if the nodes of the graph only include the adult cell types (as in Fig. 3 E).
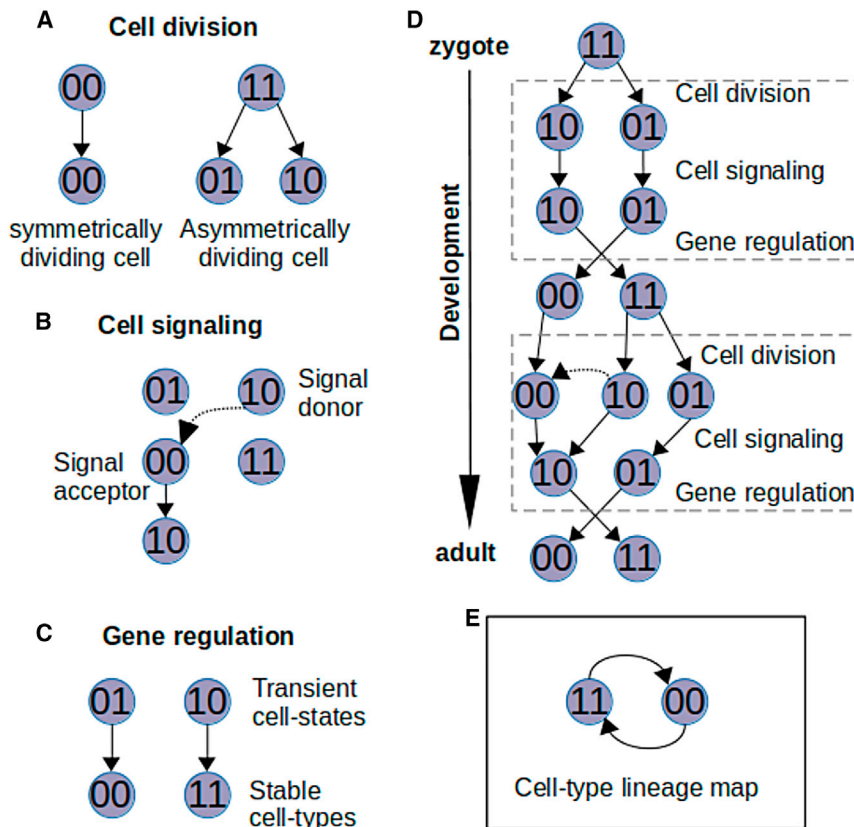3) Regeneration if the nodes of the graph include cell types produced during regeneration.

FIGURE 3 Generative model of biological development. Blue circles represent cell states, and numbers written inside them represent the identity of cell states; 0 represents the absence, and 1 represents the presence of a cell-state determinant. Solid arrows represent change in the cell state, and dashed arrows represent the exchange of signaling molecules. (A)–(C) represent the regulatory architecture of an organism with $N = 2$ determinants. (A) Asymmetric cell division: cell types in the model can produce daughter cells that are not identical. (B) Cell signaling: certain cell-state determinants can act as signals and are secreted by donor cells and received by specific acceptor cells. In this example, the first determinant is a signal. The state of the acceptor cell reflects signal reception by switching the state of signal determinant to "1." (C) Gene regulation: certain cell states are stable cell types, and others are transient cell states that map to the stable cell types. (D) Scheme of development in the model: the zygote undergoes repeated rounds of asymmetric cell division, cell-cell signaling, and gene regulation according to the rules outlined in (A)–(C). The grey dashed boxes indicate one repeat of cell division, cell signaling and gene regulation steps. The process is iterated until the resulting set of cell types repeats itself; this set of cell types forms the adult. (E) CTLM of adult homeostasis: in this example, the two cell types constitute the adult produced by the developmental program sketched in (D). The arrows represent differentiation. To see the figure in color, go online.

The model is "generative" because we can randomly draw a large number of developmental programs and their corresponding CTLMs, typically a few millions, and statistically analyze their features. Even such a large set remains minute compared with the overall number of potential developmental programs. However, the sampling of the set provides reliable statistics and thereby overcomes the "curse of dimensionality" (37) that follows from Cantor's theorem.

Insights from the generative model

In (10), we use the model described above to generate developmental programs in an unbiased manner and look at the homeostatic adult CTLMs it produces. This is the distribution of cell-lineage map topologies we should expect for instance, in multicellular lineages in which developmental plasticity has erased all traces of the mode of cell type invention.

The CTLMs generated by the model are classified according to their graph topologies: unicellular, (Appendix) cyclic, chain, tree, and (Appendix) DAGs (Fig. 4 *A*). Here, DAG specifically refers to acyclic graphs that are not tree-like and possess edges cross-linking different branches. Because differentiation in multicellular organisms is generally assumed to be irreversible, acyclic CTLMs, chain, tree, and DAG are more likely to be biologically relevant. Counter to expectations, our results indicate that tree-like graphs were extremely rare (1% of all lineage maps). The most common acyclic lineage maps were simple two-node chains, which resemble *Volvox* (Fig. 1 *C*). Among the more complex acyclic graphs with more nodes, DAGs were the most common (as is the case for hydra in Fig. 1 *D*). Thus, our results suggest that CTLMs representing the adult stage of complex multicellular organisms are very likely to be DAGs.

In reality, we expect that not all traces of the mode of cell type evolution have been wiped out due to developmental plasticity. That is, some biases and patterns should persist in the regulatory architecture of development and thereby could leave a signature on the cell-lineage map topologies. In principle, such patterned developmental programs can be described by the generative framework, which can be used to also survey and predict the distribution of CTLM topologies for distinct modes of cell type invention. Such a study could be used for comparison with experimentally obtained cell-lineage maps, allowing us to assess the modes of cell type invention prevalent in different multicellular lineages.

The topologies of generated CTLMs also indicate the ability of organisms to regenerate. The regenerative capacity of organisms was computed by separating single-cell types from the adult and testing whether they were able to regenerate the complete adult. In other words, we tested whether the adult organisms of the model contained pluripotent cells. Although we found organisms with chain-like and DAG-like lineage maps to be remarkably regenerative, those with tree-like CTLMs turned out to be the least regenerative (Fig. 4 *B*). This illustrates that topologies of CTLMs—in addition to being a summary of developmental events and indicative of the evolution of developmental programs—can also hold functional information, in this case, the ability to regenerate.

CONCLUSIONS

The evolution of multicellular complexity is synonymous with the evolution of cell types in multicellular lineages. Adding new cell types to an organism involves rewiring of its GRN. We have outlined here three modes through which such rewiring of the regulatory architecture can lead to cell type invention.
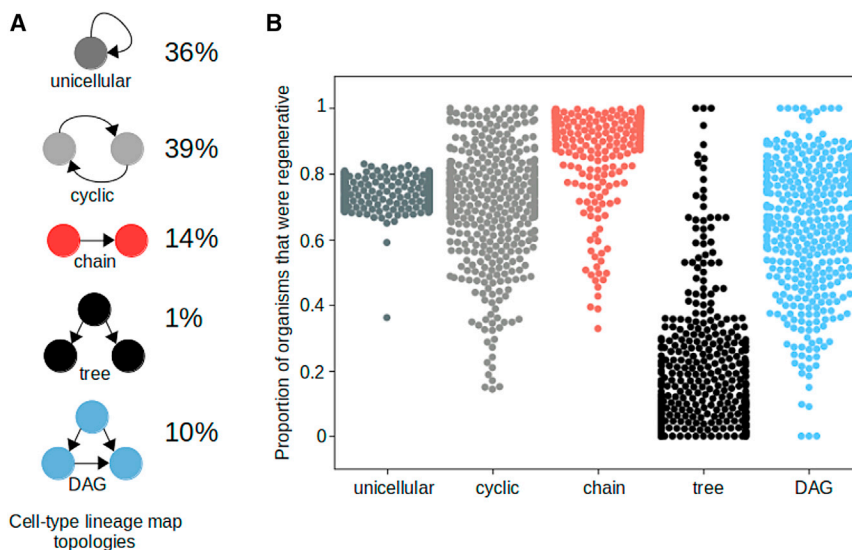

FIGURE 4 Prevalence of topologies and their regenerative capacities. (*A*) Graph topologies of CTLMs. The numbers beside the graphs indicate their prevalence in our data. (*B*) Distribution of regenerative capacities of organisms with different CTLM topologies is represented by the distribution of points in the swarms. To see the figure in color, go online.

The GRN of an organism prescribes not only its cell types but also developmental routes to reach these cell types. Thus, the prevalent modes of cell type invention are also likely to be reflected in its CTLMs. Mathematical models of development, like ours, can be used to map the regulatory architecture of development to CTLMs. These can be useful in anticipating the properties of CTLMs resulting from developmental programs that have been sculpted by different modes of cell type evolution.

We used the model to survey millions of developmental programs, unbiased toward any particular mode of cell type evolution. The survey produced a characteristic distribution of cell-lineage map topologies: tree-like lineage maps were extremely rare, and complex multicellular lineage maps were more likely to be represented by DAGs. We suggest that a combination of modeling approaches that predict topologies of cell-lineage maps, and surveys of cell-lineage maps of real organisms could uncover the patterns of developmental evolution in the various multicellular lineages. These results demonstrate that minimal coarse-grained models of development could serve as a complementary approach to detailed molecular models: while detailed models describe the implementation of development in a particular organism, coarse-grained models can be used to scan a huge space of programs and produce general conclusions about developmental processes. Importantly, mathematical models allow simplified conceptions of complicated biological processes and produce experimentally testable predictions about core features of the functioning, origin, and evolution of these processes.

APPENDIX

The following are formal definitions of a few central terms, which are mostly familiar on an intuitive level.

1) Multicellular complexity is the concept that is roughly equivalent to the number of distinct cell types an organism possesses. Simple multicellular organisms possess two cell types, usually a germ cell and a somatic cell, whereas complex multicellular organisms have elaborate bodies and can contain hundreds of cell types.
2) Developmental program is the set of instructions that can be 1) used by the single-celled zygote to form a complete adult, 2) for the maintenance of adult body, or 3) for regeneration of the body postinjury.
3) CTLM is a graph in which the nodes are the different cell types of an organism, and a directed edge between two nodes indicates that one cell type differentiates into the other during the course of the organism's development.
4) Cellular context is the set of cell types that co-occurs with a given cell type in the developing body. Cells interact with their cellular context through signaling, and this interaction regulates their differentiation trajectory.
5) Graph topology is the manner in which the nodes and edges of a graph are arranged. The topological properties of graphs include the length of its shortest path, the degree distribution of its nodes, presence of node clusters, loops, etc. Certain graph topologies are particularly recognizable, such as cycles, chains, and trees.
6) Cyclic graph is a graph that contains at least one cyclic path, i.e., a path that begins and ends at the same node. Presence of cycles in CTLMs indicate programmed reversibility of cellular differentiation.
7) Chain is a connected linear graph that contains exactly one directed path between a starting node and end node. A two-node chain CTLM represents simple multicellular organisms such as volvox.
8) Tree is any connected acyclic graph with n nodes and $n - 1$ edges, where n is an integer greater than 2. Such graphs are characterized by paths that look like branches of a tree. In CTLMs with tree-like topologies, there is exactly one developmental route that can access any cell type.
9) DAG is a connected graph that contains no cycles. Although chains and trees are also DAGs, in general, DAGs contain edges that link its different branches. In a CTLM, these links represent multiple developmental routes that converge on the same cell type.
10) Clonal multicellularity is the form of multicellularity in which the body of the organism grows by the repeated division of a single initial cell. The body is necessarily composed of clonally related cells, for example, in animals and land plants.
11) Aggregative multicellularity is the form of multicellularity in which the body of the organism grows by the aggregation of cells of the same species. Therefore, the organism need not be composed of clonally related cells. For example, spores of dictyostelids are aggregatively multicellular.
12) GRNs represent genetic interactions that regulate the activity of genes or gene products. Mathematically, GRNs can be represented as graphs in which the nodes represent genes, and the presence of an edge between two nodes indicates that one of the genes regulates the activity of the other.

AUTHOR CONTRIBUTIONS

Authors designed research, wrote, and edited the manuscript.

ACKNOWLEDGMENTS

REFERENCES

1. Meyerowitz, E. M. 2002. Plants compared to animals: the broadest comparative study of development. *Science.* 295:1482–1485.

2. Knoll, A. H. 2011. The multiple origins of complex multicellularity. *Annu. Rev. Earth Planet. Sci.* 39:217–239.

3. Arendt, D., J. M. Musser, …, G. P. Wagner. 2016. The origin and evolution of cell types. *Nat. Rev. Genet.* 17:744–757.

4. Matt, G., and J. Umen. 2016. Volvox: a simple algal model for embryogenesis, morphogenesis and cellular differentiation. *Dev. Biol.* 419:99–113.

5. Cantor, G. 1891. Ueber eine elementare Frage der Mannigfaltigkeitslehre. *Jahresbericht der Deutschen Mathematiker-Vereinigung.* 1:72–78.

6. Albert, R., and H. G. Othmer. 2003. The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster. J. Theor. Biol.* 223:1–18.

7. Milo, R., P. Jorgensen, …, M. Springer. 2010. BioNumbers–the database of key numbers in molecular and cell biology. *Nucleic Acids Res.* 38:D750–D753.

8. Barresi, M., and S. Gilbert. 2020. Developmental Biology, Twelfth Edition. Oxford University Press, New York.

9. Yuan, M., X. Yang, …, J.-R. Yang. 2020. Alignment of cell lineage trees elucidates genetic programs for the development and evolution of cell types. *iScience.* 23:101273.

10. Mani, S., and T. Tlusty. 2021. A comprehensive survey of developmental programs reveals a dearth of tree-like lineage graphs and ubiquitous regeneration. *BMC Biol.* 19:111. https://doi.org/10.1186/s12915-021-01013-4.

11. Imran Alsous, J., P. Villoutreix, …, S. Y. Shvartsman. 2017. Collective growth in a small cell network. *Curr. Biol.* 27:2670–2676.e4.

12. Sebé-Pedrós, A., B. M. Degnan, and I. Ruiz-Trillo. 2017. The origin of Metazoa: a unicellular perspective. *Nat. Rev. Genet.* 18:498–512.

13. Siebert, S., J. A. Farrell, …, C. E. Juliano. 2019. Stem cell differentiation trajectories in *Hydra* resolved at single-cell resolution. *Science.* 365:eaav9314.

14. Plass, M., J. Solana, …, N. Rajewsky. 2018. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science.* 360:eaaq1723.

15. Sogabe, S., N. Nakanishi, and B. M. Degnan. 2016. The ontogeny of choanocyte chambers during metamorphosis in the demosponge Amphimedon queenslandica. *Evodevo.* 7:6.

16. Siebert, S., F. Anton-Erxleben, and T. C. Bosch. 2008. Cell type complexity in the basal metazoan Hydra is maintained by both stem cell based mechanisms and transdifferentiation. *Dev. Biol.* 313:13–24.

17. Matsumoto, Y., S. Piraino, and M. P. Miglietta. 2019. Transcriptome characterization of reverse development in *Turritopsis dohrnii* (Hydrozoa, Cnidaria). *G3 (Bethesda).* 9:4127–4138.

18. Brunet, T., A. H. Fischer, …, D. Arendt. 2016. The evolutionary origin of bilaterian smooth and striated myocytes. *eLife.* 5:e19607.

19. Tritschler, S., M. Büttner, …, F. J. Theis. 2019. Concepts and limitations for learning developmental trajectories from single cell genomics. *Development.* 146:dev170506.

20. Lähnemann, D., J. Köster, …, A. Schönhuth. 2020. Eleven grand challenges in single-cell data science. *Genome Biol.* 21:31.

21. Girard, L. R., T. J. Fiedler, …, M. Chalfie. 2007. WormBook: the online review of *Caenorhabditis elegans* biology. *Nucleic Acids Res.* 35:D472–D475.

22. Pellin, D., M. Loperfido, …, L. Biasco. 2019. A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. *Nat. Commun.* 10:2395.

23. Guo, G., M. Huss, …, P. Robson. 2010. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Dev. Cell.* 18:675–685.

24. Treutlein, B., D. G. Brownfield, …, S. R. Quake. 2014. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature.* 509:371–375.

25. Wagner, D. E., C. Weinreb, …, A. M. Klein. 2018. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science.* 360:981–987.

26. Yanni, D., S. Jacobeen, …, P. J. Yunker. 2020. Topological constraints in early multicellularity favor reproductive division of labor. *eLife.* 9:e54348.

27. Schmidt, S. T., S. M. Zimmerman, …, S. R. Quake. 2017. Quantitative analysis of synthetic cell lineage tracing using nuclease barcoding. *ACS Synth. Biol.* 6:936–942.

28. Kebschull, J. M., and A. M. Zador. 2018. Cellular barcoding: lineage tracing, screening and beyond. *Nat. Methods.* 15:871–879.

29. Huson, D. H., and D. Bryant. 2006. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* 23:254–267.

30. Varahan, S., A. Walvekar, …, S. Laxman. 2019. Metabolic constraints drive self-organization of specialized cell groups. *eLife.* 8:e46735.

31. König, S. G., and A. M. Nedelcu. 2020. The genetic basis for the evolution of soma: mechanistic evidence for the co-option of a stress-induced gene into a developmental master regulator. *Proc. Biol. Sci.* 287:20201414.

32. Kin, K., and P. Schaap. 2021. Evolution of multicellular complexity in the dictyostelid social amoebas. *Genes (Basel).* 12:487.

33. Erkenbrack, E. M., J. D. Maziarz, …, G. P. Wagner. 2018. The mammalian decidual cell evolved from a cellular stress response. *PLoS Biol.* 16:e2005594.

34. Gershenson, C. 2004. Introduction to random Boolean networks. *arXiv preprint nlin/0408006.*

35. Rand, D. A., A. Raju, M. Sáez, F. Corson, E. D. Siggia…, 2021. Geometry of gene regulatory dynamics. *Proc Nat Acad Sci USA.* 118 (38). e2109729118. https://doi.org/10.1073/pnas.2109729118.

36. Doherty, C. A., R. Diegmiller, …, S. Y. Shvartsman. 2021. Coupled oscillators coordinate collective germline growth. *Dev. Cell.* 56:860–870.e8.

37. Eckmann, J.-P., and T. Tlusty. 2021. Dimensional reduction in complex living systems: where, why, and how. *BioEssays.* 43:e2100062. https://doi.org/10.1002/bies.202100062.